Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

a) Day of week did not have any effect on the rentals
b) Weather condition (storms) had an inverse effect on rentals(with high -ve coeff)
c) summer and fall (good weather) had better rentals than winter (bad weather)

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

When having n dummies for n levels of categorical variables, there is a redundancy. One of those columns can be represented by a combination of the rest of the columns. Hence it's a good idea to have one column (dummies) less (n-1) than the total number of categorical levels. And drop_first facilitates that, by dropping the first level.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

“year” has the highest positive correlation with the target variable. But given that it only has values of 0,1, if we want to ignore “year”. Then, “windspeed” has maximum correlation (-ve) with target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

a) Normally distributed error terms
b) Normal distribution of residuals
c) Zero mean
d) Independent (Multicollinearity is eliminated. All predictors have VIF less than 5)

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

a) Season (fall will likely have more rentals)
b) Snow storms will greatly reduce rentals
c) Windspeed has high negative correlation with rentals
d) 2019 yr like has more rentals

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

a) Analyze the data and drop unnecessary columns (unique, duplicate etc.)
b) Visualize the data (pairplot, heatmap, boxplot) to identify the collinearity with respect to target variable
c) Convert categorical variables to dummies
d) Split to train/test data
e) Scale (minmax, standardized) numeric variables
f) Outlier treatment
g) Reduce number of features (Recursive feature elimination or manual process)
h) Build linear regression model
i) Check VIF to make sure we only have features which have less than 5 (or sometimes 10, based on business needs). Eliminate one feature at a time and keep doing linear regression every time a feature is removed to make sure model is still acceptable (r2, p-value, p(f-stat))
j) Plot error (ytrain, ypred) to verify errors are normally distributed with mean zero.

k) Predict target for test dataset.

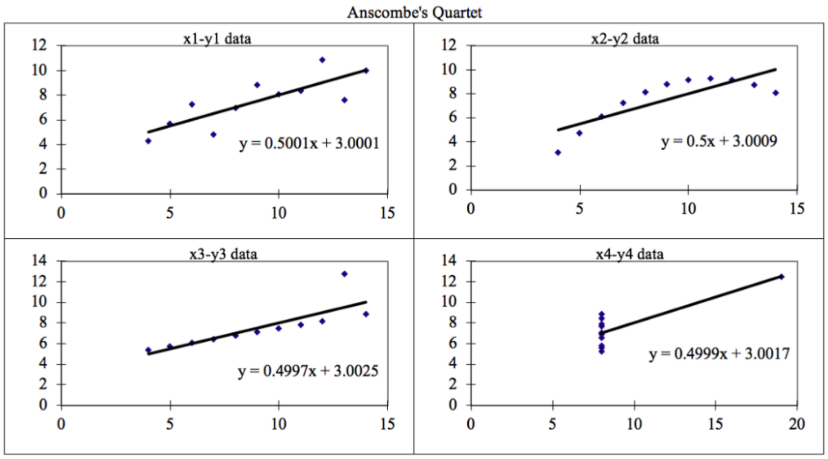
l) Compute and plot residuals and make sure they are also normally distributed.

m) Check r2-score against test data and make sure the r2-score is close to the one we got for the training set.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Example:



(source: wikipedia)

3. What is Pearson’s R? (3 marks)

Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

Source: https://statistics.laerd.com/

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. (source: geeksforgeeks). If scaling is not done, then the beta coefficients are not interpretable across the features.

Min-MaxScaling:

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1.

Standardardized Scaling:

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
(3 marks)

If there is perfect correlation between variables, then VIF is infinite.
A large value of VIF indicates that there is a correlation between the variables

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)

Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

In linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

source: https://data.library.virginia.edu/